Machine Learning applied to Rule-Based Machine Translation

Rios, Annette ; Göhring, Anne

Abstract: Lexical and morphological ambiguities present a serious challenge in rule-based machine translation (RBMT). This chapter describes an approach to resolve morphologically ambiguous verb forms if a rule-based decision is not possible due to parsing or tagging errors. The rule-based core system has a set of rules to decide, based on context information, which verb form should be generated in the target language. However, if the parse tree is not correct, part of the context information might be missing and the rules cannot make a safe decision. In this case, we use a classifier to assign a verb form. We tested the classifier on a set of four texts, increasing the correct verb forms in the translation from 78.68

DOI: https://doi.org/10.1007/978-3-319-21311-8

Machine Learning applied to Rule-Based Machine Translation

Annette Rios and Anne Göhring

Abstract Lexical and morphological ambiguities present a serious challenge in rule-based machine translation (RBMT). This chapter describes an approach to resolve morphologically ambiguous verb forms if a rule-based decision is not possible due to parsing or tagging errors. The rule-based core system has a set of rules to decide, based on context information, which verb form should be generated in the target language. However, if the parse tree is not correct, part of the context information might be missing and the rules cannot make a safe decision. In this case, we use a classifier to assign a verb form. We tested the classifier on a set of four texts, increasing the correct verb forms in the translation from 78.68%, with the purely rule-based disambiguation, to 95.11% with the hybrid approach.

1 Introduction

The term hybrid machine translation refers to any combination of statistical MT with rule-based MT (España-Bonet et al 2011) or example-based MT (Smith and Clark 2009), or a mixture of all three approaches (Alegria et al 2008).
A statistical translation system may be improved by rule-based pre-editing, such as reordering, or by the addition of linguistic features, for instance through a morphological analysis of the words in the source sentence. Furthermore, statistical methods may enhance a rule-based system on different levels: A common type of hybrid systems uses statistical ranking of translation alternatives of one rule-based system (Oepen et al 2007) or of several rule-based systems (Eisele et al 2008). Sawaf (2010) outlines yet another hybrid approach for the 'translation' of different Arabic dialect into the normalized Modern Standard Arabic: A rule-based system handles

Annette Rios
Institute of Computational Linguistics, University of Zurich, e-mail: rios@cl.uzh.ch

Anne Göhring
Institute of Computational Linguistics, University of Zuriche-mail: goehring@cl.uzh.ch

rare word combinations or phrasal structures, whereas statistical methods are used in situations where word combinations and phrasal structures occur frequently enough to estimate reliable statistics.

The hybrid architecture that we describe in this chapter consists of a rule-based core system that uses statistical modules for certain disambiguation tasks. As for the language pair in question, Spanish-Quechua, the amount of parallel text is too small to train a statistical MT system, we use a rule-based approach that relies on linguistic information and transfer rules. Nevertheless, certain ambiguities are extremely difficult to handle in a purely rule-based setting. For instance, if a word has more than one translation in the dictionary, a device for lexical selection is necessary in order to output the correct translation in the given context. This procedure presents a great challenge for a rule-based architecture, as it is not feasible to cover all possible contexts with rules. A possible solution can be to use a statistical MT system to fill in the template of the target sentence generated by the rule-based system (España-Bonet et al 2011), (Hunsicker et al 2012). If no MT system is available, another option is to use a machine learning approach, e.g. sequence labeling (Rudnick and Gasser 2013) or to generate all possible translations and use a statistical language model to score the alternatives (Melero et al 2007).

Words may not only have different lexical translations, there can also be morphological ambiguities: a word may have more than one translation with the same lemma, but different morphology. A set of rules that match the context of the verb decides which verb form should be generated in the target language. However, due to parsing or tagging errors, these rules might not be applicable in all cases. In this chapter, we will present an approach to disambiguate such morphological ambiguities with machine learning.

2 SQUOIA Spanish to Quechua MT System

As part of our research project SQUOIA[1], we have implemented a mostly rule-based machine translation system that translates text from Spanish to Quechua. The system uses a classical transfer approach, where several modules are joined in a processing chain: each module relies on the output of the previous module for further processing, see Fig. 1 for an overview. One of the most difficult parts during the translation is the disambiguation of subordinated Spanish verbs in order to generate the correct Quechua forms, as the grammatical features encoded in verbs differ considerably between these two languages. To a certain extent, subordinated verb forms can easily be disambiguated with a set of rules, but this strategy is not practical in all cases. In this chapter, we will present an approach that uses machine learning to resolve verb forms in contexts that cannot be safely disambiguated by rules.

There are two kinds of subordinated clauses that we need to disambiguate: clauses with a verbal head (complement clauses, final clauses, etc.) and clauses with

[1] `http://tiny.uzh.ch/2Q`

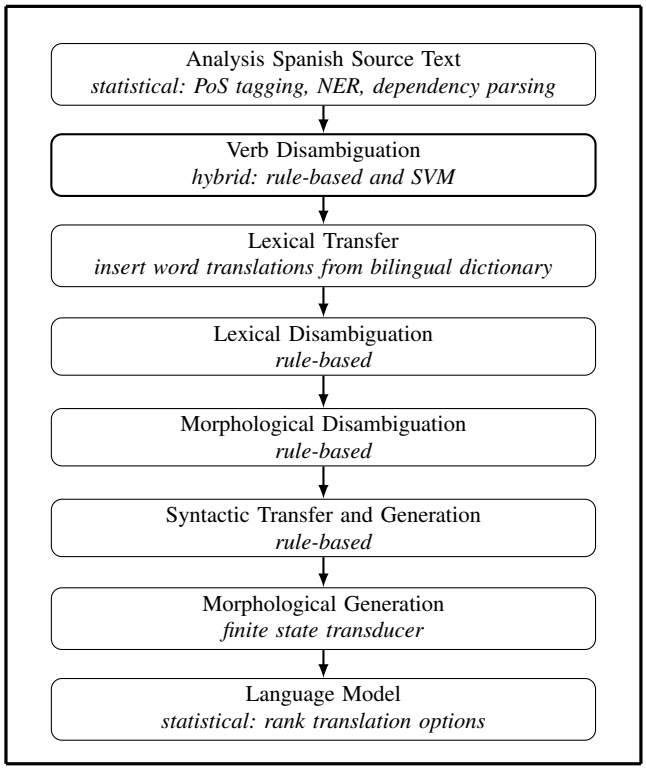

Fig. 1 SQUOIA Translation Pipeline Spanish-Quechua

a nominal head (relative clauses). In both cases, we use a set of rules to determine which Quechua verb form should be generated in the given context. For relative clauses, we have to rely on semantic information about the head and the subcategorization frames of the verb, whereas for other subordinated clauses, we need the conjunction and the semantics of the main verb to determine the correct Quechua verb form. In a real application scenario however, we might not have access to all the information we need to make a rule-based decision, due to tagging or parsing errors.

In the case of relative clauses, it is important to note that the syntactic structure alone does not always allow for a safe decision, as Spanish relative clauses can be highly ambiguous. In this case, the rule-based module guesses the correct form based on semantic information.

In a previous experiment, we extracted context information about subordinated clauses with verbal heads from two treebanks and trained different classifiers on this data (Rios and Göhring 2013). In this first setup, we used the lemmas of the main and the subordinated verb as attributes. However, as the decision relies on lemmas, we might have a problem with sparse data, as the classifier has only information about the lemmas seen in training. Therefore, we will present an alternative

approach in this chapter that relies on semantic information about verbs[2] and verb frames[3]. In the previous setting, Naïve Bayes achieved the best results, with 81% in 10-fold cross-validation and 84% on a separate test set. With the new set of features, the independence assumption may not always be true anymore. As a consequence, Naïve Bayes is no longer an option, and so we decided to use libsvm (Chang and Lin 2011) instead. We were able to increase the accuracy with this new approach to 92% in cross-validation and 86% on the same test set.

3 Subordinated Quechua Verb Forms

Subordinated clauses in Quechua are often non-finite, nominal forms. There are several nominalizing suffixes that are used for different clause types that will be illustrated in more detail in this section.

3.1 Switch-Reference

A common type of subordination in Quechua is the so-called switch-reference: the subordinated, non-finite verb bears a suffix that indicates whether its subject is the same as in the main clause or not. If the subject in the subordinated clause is different, the non-finite verb bears a possessive suffix that indicates the subject person. Consider the following examples:[4]

[2] extracted from the Spanish part of Multilingual Central Repository 3.0 (Gonzalez-Agirre et al 2012).

[3] extracted from the AnCora verb lexicon (Taulé et al 2008).

[4] Abbreviations used:

Acc: accusative	Add: additive ('too,also')
Ag: agentive	Ben: benefactive ('for')
Con: connective ('and')	Dir: directional
DirE: direct evidentiality	DS: different subject
Gen: genitive	Imp: imperative
Inch: inchoative	Loc: locative
Neg: negation	Obl: obligative
Perf: perfect	Poss: possessive
Prog: progressive	Pst: past
Rflx: reflexive	Sg: singular
SS: same subject	Top: topic

(1) Same subject: *Mikhuspa hamuni.*

Mikhu	***-spa***	*hamu*	*-ni.*
eat	**-SS**	come	-1.Sg

"When I finished eating, I'll come."
(lit. "My eating, I come.")

(2) Different subject: *Mikhuchkaptiy pasakura.*

Mikhu	*-chka*	***-pti***	*-y*	*pasa*	*-ku*	*-ra*	*-ø.*
eat	-Prog	**-DS**	-1.Sg.Poss	leave	-Rflx	-Pst	-3.Sg

"While I was eating, he left."
(lit. "my being-eating, he left.")

(Dedenbach-Salazar Sáenz et al 2002:168)

In the source language, Spanish, subordinated verbs are usually finite. An overt subject is not necessary, as personal pronouns are used only for emphasis ("pro-drop"). In order to generate the correct verb form, we need to find the subject of the subordinated verb and compare it to the main verb. For this reason, we included a module that performs coreference resolution on subjects. So far, the procedure is based on the simple assumption that an elided subject is coreferential with the previous explicit subject, if this subject agrees in number and person with the current verb. However, some exceptions have to be considered, e.g. the subject of a verb in direct speech is not a good antecedent.

3.2 Other Types of Subordination

Generally, the relation of the subordinated clause to the main clause is expressed through different conjunctions in Spanish. In Quechua, on the other hand, a specific verb form in combination with a case suffix indicates the type of subordination. For instance, Spanish *para que* - "in order to" has to be translated with a nominal verb form with the suffix *-na* ('obligative') and the case suffix *-paq* (usually called benefactive, "for"):

(3) *Ventanata kichay wayraq haykurimunanpaq.*

Ventana	*-ta*	*kicha*	*-y*	*wayra*	*-q*	*hayku*	*-ri*	*-mu*	***-na***	*-n*
window	-Acc	open	-2.Sg.Imp	wind	-Gen	enter	-Inch	-Dir	**-Obl**	-3.Sg.Poss

-paq.
-Ben

"Open the window, so the air comes in."
(lit. "Open the window for his entering of the wind")

(Cusihuamán 2001:210)

Finite verb forms are also possible in subordinated clauses; in this case, the relation of the subordinated and the main clause is indicated through a "linker". A linker often consists of a demonstrative pronoun combined with case suffixes or so-called independent suffixes; these are special suffixes that can be attached to any word class and their position is usually at the end of the suffix sequence. The functions of the independent suffixes include data source, polar question marking and topic or contrast, amongst others (Adelaar and Muysken 2004:209). In combination with demonstrative pronouns, the independent suffixes are used for linking clauses, similar to Spanish or English conjunctions. For instance, the combination of demonstrative *chay* - "this" with the topic marker *-qa*, *chayqa*, is used in the sense of "if, in case that":

(4) *Munanki **chayqa**, Arekipatapis rinki makinapi.*

Muna	*-nki*	***chay***	***-qa***,	*Arekipa*	*-ta*	*-pis*	*ri*	*-nki*	*makina*	*-pi.*
want	-2.Sg	**this**	**-Top**	Arequipa	-Acc	-Add	go	-2.Sg	machine	-Loc

"If you like, you can also go to Arequipa by train (machine)."

(Cusihuamán 2001:264)

Indirect speech in the Spanish source text is a special case, as the Quechua equivalence of indirect speech is direct speech. The conversion from indirect to direct speech is not trivial, because coreference resolution for the subject is required: if the subject of the main verb is the same as the subject of the indirect speech clause, the verb has to be generated as first person form in direct speech. Consider this English example:

(5) "John said he wanted to go fishing."

a. *if John = he:* "I want to go fishing", John said.

b. *if John ≠ he:* "He wants to go fishing", John said.

In this case, we naively consider both subjects as being equal and mark the direct speech Quechua verb as a first person form, as the current rule-based approach is not good enough to distinguish these two cases. However, we plan to integrate a statistical means for coreference resolution in order to make better decisions as to which form should be generated.

Furthermore, the form of the subordinated verb may also depend on the semantics of the main verb, e.g. complement clauses of control verbs usually require *-na* ('obligative'), whereas with other verbs, the nominalizer *-sqa* ('nominal perfect') is used[5]:

[5] Double marking of negation in (6.b): *ama:* negation particle in imperative clauses ('don't'), *-chu*: negation suffix, attached to the constitutent in focus

(6) a. *Ri* ***-na*** *-yki* *-ta* *muna* *-ni.*
go **-Obl** -2.Sg.Poss -Acc want -1.Sg

"I want you to leave."
(lit. "I want your going.")

b. *Ama* *-n* *chay* *yacha* ***-sqa*** *-yki* *-ta* *qunqa* *-nki* *-chu.*
don't -DirE this know **-Perf** -2.Sg.Poss -Acc forget -2.Sg -Neg

"Don't forget what you learned."
(lit. "Don't forget those your learned-ones.")

(Cusihuamán 2001:125)

For all of these cases, the translation system has a set of rules to match the given context, so that the correct form can be assigned to each verb.

4 Verb Form Disambiguation with Machine Learning

4.1 Training Data

In order to generate the correct Quechua verb form in a subordinated clause, we need to extract the following information from the Spanish source sentence:

- semantics of the main verb
- the conjunction
- tense and mood of the subordinated verb (in some cases needed to distinguish between 'obligative' *-na* and 'perfect' *-sqa*)

Based on these features, the rule-based verb disambiguation module of the translation system assigns the Quechua verb form. Given a correct dependency tree, this rule-based approach achieves a high precision, but it is bound to fail if the parse tree is erroneous. In order to obtain instances of main and subordinated clauses for training a classifier, we pre-translated two manually annotated dependency treebanks: the Spanish AnCora dependency treebank[6] (Taulé et al 2008) and the IULA Spanish LSP Treebank[7] (Marimon et al 2012). As these are correctly annotated, the rule-based module can disambiguate the subordinated verbs with great reliability, and we can extract these clauses as instances for training. With this approach, we collected 8,579 instances from AnCora and 5,704 from IULA[8], which results in a total of 14,283 instances for training.

[6] `http://clic.ub.edu/corpus/en/ancora`

[7] `http://www.iula.upf.edu/recurs01_tbk_uk.htm`

[8] Note that, although IULA contains more than twice as many sentences as AnCora, the sentences in IULA are mostly short, simple sentences, without subordinated clauses.

4.2 Features

Instead of lemmas we use the semantic categories from the Spanish wordnet (Gonzalez-Agirre et al 2012) and the AnCora verb frames (Taulé et al 2008) to describe the main verb. For the subordinated verb, only tense and mood are relevant (extracted from the PoS tag in the treebank). For the conjunctions, we use the lexical forms, as there is no good way to describe them semantically. All features are binarized for training.

In our previous pipeline (Rios and Göhring 2013) we relied on the lemmas of main and subordinated verb instead of semantic and syntactic features. In this setting, Naïve Bayes achieved the best results, yet as with the new set of features, the independence assumption might not always be given, we switched to support vector machines.

4.3 Classification

We decided to use libsvm for the classification, as it provides a simple way of optimizing the parameters *c* (cost) and *g* (gamma) via grid search. Table 1 shows the accuracy of libsvm in 10-fold cross validation and on a manually annotated test set of 100 instances. This is the same test set that we used before with Naïve Bayes (Rios and Göhring 2013). For comparison, Table 1 also contains the results obtained with Naïve Bayes, once trained on the exactly same data set as libsvm, and once trained on the same data, but with verb lemmas instead of semantic and syntactic features. The results in Table 1 indicate that libsvm achieves the best accuracy, with 92.07% in cross-validation and 86% on the test set.[9]

The classification is slightly worse if only the conjunction and the subordinated verb are set, but the main verb is unknown (second line in Table 1). The third option, that the classifier has only information about the main and the subordinated verb while the conjunction is unknown, is not relevant: In case no conjunction has been found, the module assumes that the verb form in question must be either a main verb, a relative clause or a coordination. All of these options are set by rules, not by the SVM classifier.

[9] In our previous setting with Naïve Bayes, we achieved only 81% accuracy, but we had a smaller training set of only ∼7,300 instances.

[10] C-support vector classification (C-SVC) with RBF kernel parameters c (cost) and g (gamma) obtained through search grid on 10-fold cross-validation (10x cv)

Table 1 Evaluation[10]

Features	libsvm C-SVC, RBF, c=32, g=0.0078125		Naïve Bayes with semantic/ syntactic feat.		Naïve Bayes with lemmas	
	10x cv	test set	10x cv	test set	10x cv	test set
main verb, sub. verb, conjunction	**92.08**	**86**	81.47	75	84.28	78
sub. verb, conjunction	**87.97**	**81**	85.07	75	74.02	72

4.4 RBMT System with SVM Verb Disambiguation

Figure 2 illustrates how the SVM module is integrated into the translation pipeline: The rule-based verb disambiguation module tries to assign a Quechua form to all verbs in the Spanish tree. If the main verb or the conjunction is not found during this rule-based disambiguation, the verb form is marked as ambiguous and passed to the additional module for further disambiguation. This additional module checks in a first step if a given ambiguous verb form could be the actual main verb of the sentence or a relative clause that the parser attached to a non-nominal head. If this is the case, it assigns the verb form *finite* or *rel* for main or relative clauses respectively, and the disambiguation is done. Otherwise, it checks if there is a conjunction, if so, it looks for the main verb in the linear sequence of the tokens,[11] and then invokes the SVM model to assign a verb form. If there was no conjunction, the module assumes that this must be a coordination and assigns the same verb form as the preceding verb. If there is no preceding verb, this might be a tagging error, in this case the module assigns the verb form *finite*, as this is the most common form.

4.5 Evaluation

Whole verb disambiguation pipeline

We used the same four texts for the evaluation as in the previous setup (Rios and Göhring 2013):

- *La catarata de la sirena* - 'the waterfall of the siren' (Andean story)
- first two chapters of 'The Little Prince'
- article from the Peruvian newspaper 'El Diario'
- Spanish Wikipedia article about Peru

Since our previous publication (Rios and Göhring 2013), we have improved our tagger, and therefore the number of recognized verbs is slightly higher than in the version from 2013. The rule-based module disambiguates only 78.67% of all verb

[11] The first verb to the left or right that is not an auxiliary and with no conjunction or relative pronoun between them.

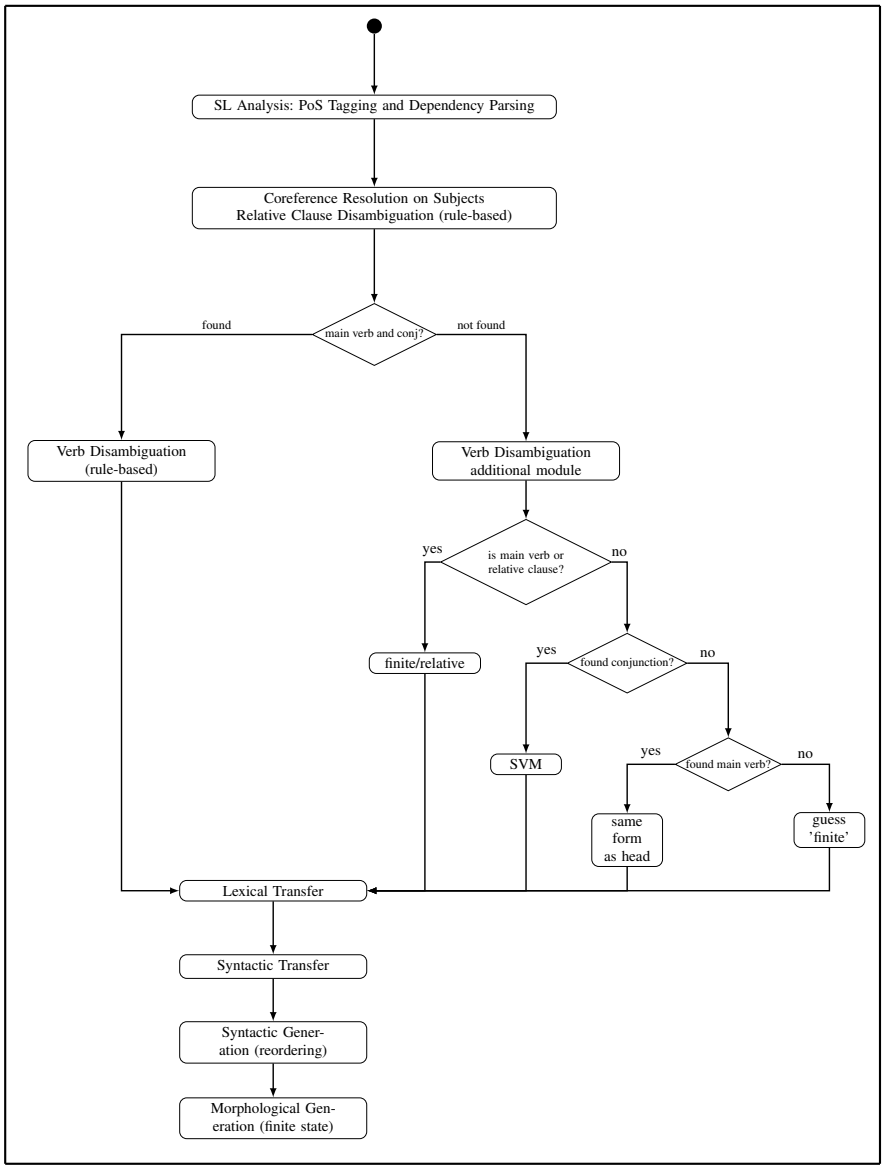

Fig. 2 SVM Module in MT Pipeline

forms correctly, as it marks many verbs as ambiguous. In the next step, the additional disambiguation module with the SVM classifier assigns a verb form to all the ambiguous forms and thus increases the proportion of correct verb forms to 95.11%. The previous module, with Naïve Bayes, achieved only 89% accuracy on these texts, see Table 2.

Table 2 Evaluation of the complete Disambiguation Pipeline

		correct	incorrect
rule based:	186	177	9
		78.67%	4%
with additional module (includes SVM) :	39	37	2
total "verb" chunks:	225	214	11
		95.11%	**4.89%**
old version, with Naïve Bayes:		89%	11%

Additional verb disambiguation module

Furthermore, we used three larger texts to test the performance of the rule-based and the SVM part of the additional verb disambiguation module. As shown in Fig. 2, the additional module relies on a set of rules to decide if the 'subordinated' verb in question is the actual main verb, a relative clause or a coordinated clause. If this is not the case, but the clause is clearly subordinated (indicated through a conjunction), the verb form is determined via SVM.

The texts that we used for this evaluation are:

- Festschrift 40th anniversary of the Peruvian-German chamber of commerce and industry (322 sentences)[12]
- Memoria 2009, Peruvian-German chamber of commerce and industry (314 sentences)[13]
- *La papa y el cambio climático* - 'potatoes and climate change', inforesources 2008 (development aid, 456 sentences)[14]

Table 3 illustrates the performance of the additional verb disambiguation module. Most of the potential ambiguous verbs (73 out of 92) are either main verbs, relative

[12] `http://www.camara-alemana.org.pe/Publicaciones/MIGEdiciones/2010MEMORIA2009.pdf`

[13] `http://www.camara-alemana.org.pe/Publicaciones/MIGEdiciones/2010MEMORIA-JAHRESBERICHT2009x.pdf`

[14] `http://www.inforesources.ch/pdf/focus08_1_s.pdf`

clauses or coordinations that had been attached to the wrong head and could therefore not be disambiguated by the rule-based module, but by the rule-based decision part of the additional verb disambiguation module. Not all ambiguous verb form candidates are actual verbs: the middle part of Table 3 shows 5 cases where nouns have been erroneously tagged as verbs. In total, the additional module assigned 79 out of 87 actual verb forms correctly, which results in 90.8% accuracy.

Table 3 Evaluation of the Additional Verb Disambiguation Module

	rule-based decision (main verb, relative clause or coordination)	SVM	total
total ambiguous verb forms	73	19	92
total correct	64	15	79
			85.87%
total wrong	9	4	13
			14.13%
total tagging errors (no verbs)	4	1	5
total disambiguated (actual verbs)	69	18	87
correct	64	15	79
			90.8%
wrong	5	3	8
			9.2%

5 Relative Clauses

5.1 Quechua Relativization

Relative clauses in Quechua are nominal forms that are either agentive or non-agentive. For non-agentive relative clauses, there are two nominalizing suffixes available: *-sqa* ('perfect') is used for actions that have been completed, whereas *-na* ('obligative') occurs in contexts where the action has not been completed or indicates an intention, obligation or purpose. Consider the following examples:

(7) a. agentive:
Wasi ruwaq runa hamuchkan.
Wasi ruwa ***-q*** *runa hamu -chka -n.*
house make **-Ag** man come -Prog -3.Sg
'The man who builds houses is coming.'
(lit. 'the house-making man is coming')

b. non-agentive:
yachasqayki llaqta
yacha ***-sqa*** *-yki* *llaqta*
live **-Perf** -2.Sg.Poss village

'the village where you live'
(Dedenbach-Salazar Sáenz et al 2002:141)

c. non-agentive:
*Qantaq, Gregorio, monta**na**y caballoyta hap'iy!*
Qan -taq, Gregorio, monta ***-na*** *-y* *caballo -y*
you -Con Gregorio ride **-Obl** -1.Sg.Poss horse -1.Sg.Poss
-ta hap'i -y!
-Acc grab -2.Imp

'And you, Gregorio, grab my riding horse!'
(lit. 'grab the horse that I will ride/intend to ride')
(Valderrama Fernández and Escalante Gutiérrez 1982)

In order to generate the correct verb form for a Quechua relative clause, it is necessary to automatically distinguish between relativization on subjects and relativization on obliques. The latter are always translated with the non-agentive forms, but relative clauses where the head noun is the subject need to be further disambiguated: If the subject is a semantic agent, the verb in the relative clause has to be rendered in the agentive form (*-q*), if the subject is not agentive, either *-sqa* or *-na* is the correct form.

Relative clauses in the source language Spanish can be very ambiguous, consider the following examples:

(8) a. agentive:
la mujer que comió la manzana
the woman REL ate the apple

'the woman who ate the apple'

b. non-agentive:
la manzana que comió la mujer
the apple REL ate the woman

'the apple that the woman ate'

The only difference between sentence (8a) and (8b) is the semantic class of the head noun: The verb *comer* - 'to eat' requires an animate, agentive subject like *mujer*. An inanimate noun like *manzana* can therefore not be the subject of *comer*. The correct translation of example (8a) uses the verb form with *-q*, whereas the verb in (8b) should be translated with *-sqa*:

(9) a. agentive:
mansana mikhu ***-q*** *warmi*
apple eat **-Ag** woman

'the woman who eats/ate the apple'

b. non-agentive:
warmi -p mikhu ***-sqa*** *-n mansana*
woman -Gen eat **-Perf** -3.Sg.Poss apple

'the apple that the woman eats/ate'

Not every Spanish relative clause is as ambiguous as the examples in (8a) and (8b). In the following cases, the head noun cannot be the subject of the relative clause, and therefore the agentive form can be discarded for the translation:

1. if the relative pronoun is preceded by a preposition (*el hombre a quien vió*),
2. if the relative pronoun is something other than *que, quien* or *cual*
3. if the verb in the relative clause is not congruent with the head noun
4. if the relative clause contains a subject noun or pronoun

Note that case 4 is not a reliable feature in the translation process, as the parser frequently labels subjects as objects and vice versa, therefore, even if the parser detected a subject in the relative clause, the following disambiguation steps will still be applied. The rule-based module uses a lexicon of Spanish verb frames (Taulé et al 2008): If the verb has only one frame, and the frame is intransitive, the head noun must be the subject. The semantic role indicated in the lexicon (agent, patient, impersonal, causer etc.) is the key to the correct translation: the Quechua verb should be rendered with the *-q* form, if the semantic role is agentive. In all other cases, the verb form in Quechua should be generated with either *-sqa* or *-na*. Whether to use the obligative or the perfect form has to be decided based on tense, aspect and mood of the Spanish verb.

If the frame retrieved from the semantic lexicon is transitive or ditransitive, the head noun is either the subject or object, but never the indirect object, as in this case the relative pronoun is preceded by the preposition *a*:

(10) indirect object as head of a relative clause:
el vecino ***a*** *quien la mujer muestra el libro*
the neighbor to REL the woman shows the book

'the neighbor, to whom the woman shows the book'

If the verb frame is transitive or ditransitive with an agentive subject, we cannot know whether the head noun is the subject or the object (see examples (8a) and (8b)). In case the verb lexicon contains more than one possible frame for a given verb, the module tries to delete all inapplicable frames with some additional context checks. If the frames cannot be reduced to one semantic role for the subject, the module takes a guess based on the semantics of the head noun. In this case, the

disambiguation module retrieves the semantic information of the head noun from a semantic noun lexicon (Marimon et al 2007): if the head noun is a likely agent (e.g. animate, human, a social group, an instrument), it assumes the agentive form, but if the head noun is an unlikely agent (e.g. an inanimate or an abstract noun, a plant) it assigns one of the non-agentive verb forms.

The basic assumption is that only nouns of certain semantic groups are plausible agents, while others are not (e.g. plants, abstract nouns, inanimates). This premise is of course not always correct, therefore we tested a machine learning approach to disambiguate relative clauses.

5.2 Relative Clause Disambiguation with Machine Learning

The disambiguation of relative clauses with machine learning differs substantially from the disambiguation of other subordinated verb forms. Section 4 illustrates how the MT system relies on a classifier to determine the Quechua verb form in cases where the analysis of the Spanish source sentence went wrong. In the experiments with relative clauses, on the other hand, we try to use a classifier to assign the correct form instead of guessing the form based on semantic information in highly ambiguous cases.

5.3 Training Data

The training material consists of automatically annotated relative clauses from the AnCora and IULA treebanks. Most relative clauses are not ambiguous: As AnCora and IULA are manually annotated, the annotation of subjects in relative clauses is reliable, as opposed to automatically parsed texts. Therefore, relative clauses that contain a subject in the treebanks are always non-agentive. Furthermore, if the verb has only intransitive frames with either agentive or non-agentive subjects, we need no further disambiguation, as we can fully rely on the semantic role of the subject given in the verb frame lexicon. The ambiguous cases in AnCora and IULA that the module had to guess were manually checked and corrected.

Note that not all relative clauses are interesting for training, as we want to use the classifier only on ambiguous forms that cannot be determined by considering only the syntactic context. With this approach, we extracted 5,018 instances from AnCora and 3,201 instances from IULA to train the classifier.

5.4 Features

In addition to the verb frames (Taulé et al 2008) and the semantic noun classes (Marimon et al 2007) used by the rule-based module, we integrated semantic information about the verb and the head noun from the Spanish wordnet (Gonzalez-Agirre et al 2012) to the classification with libsvm. The semantic noun classes of the Spanish Resource Grammar include e.g. *human*, *body part*, *plant*, *abstract noun*, etc. The classes from the Spanish wordnet overlap with these in part, but are more fine-grained for abstract nouns, they include e.g. *feeling*, *event*, *phenomenon*, *motive*, *process* and some more.

Furthermore, we included some syntactic information, or more specifically whether the relative clause contains:

- the reflexive *se*[15]
- an indirect object
- a prepositional object
- an adjunct
- the demoted subject of a passive clause
- a predicative element (in equational clauses)

Note that we did not include the presence of a subject or direct object in the relative clause as features, as we cannot safely rely on the parser for this distinction.

Furthermore, we included an additional binary feature that indicates whether the lemma of the verb in Quechua is the copula *ka-*. The reason behind this feature is that relative clauses with *ka-* use the agentive form, although the head noun is not a semantic agent. Relative clauses with the copula thus do not follow the general rule, see Example (11).

(11) *urqu* *-kuna* *-pi* *ka* *-q* *ayllu* *-kuna*
mountain -Pl -Loc be -Ag village -Pl

'mountain villages'
(lit. the villages that are in the mountains)

5.5 Evaluation

The test set consists of 106 ambiguous relative clauses extracted from Spanish Wikipedia articles about three authors: Gabriel García Márquez, Mario Vargas Llosa and Pablo Neruda.[16]

The baseline in Table 4 is the performance of the rule-based module that guesses the form based on semantic information about the head. This simple guess was

[15] The Spanish reflexive *se* is a device to render a transitive verb intransitive.

[16] `http://es.wikipedia.org/wiki/` retrieved 11.01.2014

correct in 88 out of 106 cases, which results in 83.02% accuracy. As Table 4 shows, the SVM classifier does not achieve the accuracy of the rule-based method: Even in the best setting, with all features, the classifier assigns the correct form only in 83 out of 106 cases. This results in an accuracy of 78.3%, which is slightly worse than the performance of the rule-based module.

A possible explanation is the relatively small number of training instances: although we exploited two treebanks, the training set consists of only 8,219 instances, as opposed to the 14,283 instances used to train the classifier for the subordinated verbs. Furthermore, the training material is probably not as clean as the instances used for the disambiguation of the subordinated verbs: Only the highly ambiguous (guessed) cases were manually checked, but there might as well be a number of errors in the remaining relative clauses.

Table 4 Evaluation of the SVM Classifier on Relative Clauses

	10x cv	test set
libsvm:	(C-SVC, RBF, c=8, g=0.03125)	
all features	**77.81**	**78.30**
no wordnet	75.46	75.47
no verb frames	72.89	64.15
no Resource Grammar noun classes	77.17	77.36
no syntactic features	76.10	75.47
baseline (rule-based)	–	83.02

6 Conclusions

We enhanced a purely rule-based machine translation system for the language pair Spanish-Quechua with an SVM module that predicts the form of subordinated verbs in the target language Quechua, based on information collected from the Spanish input text. The MT system has rules to match the context of the subordinated verb and assign a Quechua verb form for generation. Due to parsing and tagging errors, the information needed for this rule-based disambiguation cannot always be retrieved. In order to disambiguate these forms, we use a classifier that predicts the verb form even if all of the context information is not accessible.

We use two Spanish dependency treebanks to generate the training instances for the classifier: We let the rule-based part of the MT system assign a verb form to the subordinated clauses in the treebanks, and then extract these clauses for training. As the trees in the treebanks are annotated correctly, the rules assign the correct verb form reliably.

In a previous version of the verb disambiguation module, we used Naïve Bayes to decide the ambiguous cases, based on the lemmas of the main and the subordinated verb, as well as the conjunction. With this approach, the decision relies on lemmas,

and we might have a problem with sparse data, as the classifier has only information about the lemmas seen in training.

In order to avoid this problem, we decided to use the semantic classes from the Spanish wordnet (Gonzalez-Agirre et al 2012) and the verb frames from the AnCora Verbframe Lexicon (Taulé et al 2008) instead of lemmas. Due to the introduction of semantic classes and verb frames as features instead of lexical forms, the independence assumption may no longer be true, and therefore, we decided to use libsvm instead. Additionally, we enlarged the training set by exploiting not only AnCora (Taulé et al 2008), but also IULA (Marimon et al 2012).

Previously, Naïve Bayes achieved 81% in 10-fold cross-validation and 84% on a separate test set. We were able to increase the accuracy with the new feature set and libsvm to 92% in cross-validation and 86% on the same test set. As for now, only verb forms marked as ambiguous by the preceding rule-based module are disambiguated by the SVM module. Nevertheless, quite a large proportion of these verbs were identified as the actual main verb of the sentence. This implies that the verb that appears as head of the sentence in the parse tree should actually be a subordinated verb. In the future, we will use the SVM classifier to reassign the correct verb form to these verbs and thus increase the number of correct forms in the translation.

Furthermore, we tested if a similar approach would be suitable for the disambiguation of relative clauses, as opposed to a rule-based approach where, in ambiguous cases, the module guesses the form of the verb based on semantic information. As with the subordinated verbs, we used the rule-based module to assign a form to the relative clauses in both AnCora and IULA, and then extracted these relative clauses as instances for training, after manually checking the ambiguous forms. However, the rule-based approach still outperforms the classifier with 83.02% to 78.3%, respectively. A possible reason for the poor performance might be the relatively small number of training instances: we extracted only 8,219 relative clauses from the treebanks, as opposed to 14,283 instances of subordinated verbs.

Acknowledgments

This research is funded by the Swiss National Science Foundation under grant 100015_132219/1.

References

Adelaar WFH, Muysken P (2004) The Languages of the Andes. Cambridge Language Surveys, Cambridge University Press, Cambridge, UK

Alegria I, Casillas A, Díaz de Ilarraza A, Iguartua J, Labaka G, Lersundi M, Mayor A, Sarasola K (2008) Mixing Approaches to MT for Basque: Selecting the best

output from RBMT, EBMT and SMT. In: Proceedings of the MATMT2008 Workshop: Mixing Approaches to Machine Translation

Chang CC, Lin CJ (2011) LIBSVM: A Library for Support Vector Machines. ACM Transactions on Intelligent Systems and Technology 2(3):27:1–27:27

Cusihuamán AG (2001) Gramática Quechua: Cuzco-Collao, 2nd edn. Serie Saber Andino, Ministerio de Educación, Lima, Peru

Dedenbach-Salazar Sáenz S, von Gleich U, Hartmann R, Masson P, Soto Ruiz C (2002) Rimaykullayki - Unterrichtsmaterialien zum Quechua Ayacuchano, 4th edn. Dietrich Reimer Verlag GmbH, Berlin, Germany

Eisele A, Federmann C, Uszkoreit H, Saint-Amand H, Kay M, Jellinghaus M, Hunsicker S, Herrmann T, Chen Y (2008) Hybrid Machine Translation Architectures within and beyond the EuroMatrix project. In: Proceedings of the European Machine Translation Conference EAMT, European Association for Machine Translation, pp 27–34

España-Bonet C, Labaka G, Díaz de Ilarraza A, Màrquez L, Sarasola K (2011) Hybrid Machine Translation Guided by a Rule-Based System. In: Proceedings of the 13th Machine Translation Summit, Xiamen, China, pp 554–561

Gonzalez-Agirre A, Laparra E, Rigau G (2012) Multilingual Central Repository version 3.0: upgrading a very large lexical knowledge base. In: Proceedings of the Sixth International Global WordNet Conference (GWC'12), Matsue, Japan

Hunsicker S, Chen Y, Federmann C (2012) Machine Learning for Hybrid Machine Translation. In: Proceedings of the Seventh Workshop on Statistical Machine Translation, Montreal, Canada, pp 312–316

Marimon M, Seghezzi N, Bel N (2007) An Open-source Lexicon for Spanish. Procesamiento del Lenguaje Natural 39:131–137

Marimon M, Fisas B, Bel N, Arias B, Vázquez S, Vivaldi J, Torner S, Villegas M, Lorente M (2012) The IULA Treebank. In: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey

Melero M, Oliver A, Badia T, Suñol T (2007) Dealing with Bilingual Divergences in MT using Target Language N-gram Models. In: Proceedings of the METIS-II Workshop: New Approaches to Machine Translation, Leuven, Belgium, pp 19–26

Oepen S, Velldal E, Lønning JT, Meurer P, Rosén V, Flickinger D (2007) Towards Hybrid Quality-Oriented Machine Translation. On Linguistics and Probabilities in MT. In: Proceedings of Theoretical and Methodological Issues in Machine Translation, Skövde, Sweden

Rios A, Göhring A (2013) Machine Learning Disambiguation of Quechua Verb Morphology. In: Proceedings of the Second Workshop on Hybrid Approaches to Translation, Sofia, Bulgaria, pp 13–18

Rudnick A, Gasser M (2013) Lexical Selection for Hybrid MT with Sequence Labeling. In: Proceedings of the Second Workshop on Hybrid Approaches to Translation, Sofia, Bulgaria, pp 102–108

Sawaf H (2010) Arabic Dialect Handling in Hybrid Machine Translation. In: Proceedings of the 9th Conference of the Association for Machine Translation in the Americas

Smith J, Clark S (2009) EBMT for SMT: A New EBMT-SMT Hybrid. In: Proceedings of the 3rd Workshop on ExampleBased Machine Translation, pp 3–10

Taulé M, Martí MA, Recasens M (2008) AnCora: Multilevel Annotated Corpora for Catalan and Spanish. In: Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), Marrakech, Marroco

Valderrama Fernández R, Escalante Gutiérrez C (1982) Gregorio Condori Mamani: autobiografía. Centro Bartolomé de las Casas, Cuzco, Peru